

Combining Clustering techniques and Formal Concept Analysis to characterize Interestingness Measures

Dhouha GRISSA¹, Sylvie GUILLAUME², and Engelbert MEPHU NGUIFO¹

LIMOS, UMR 6158 CNRS, Blaise Pascal¹ and Auvergne² University,
Complexe scientifique des Cézeaux, 63177 Aubière cedex, France
{dgrissa,sylvie.guillaume,mephu}@isima.fr
<http://www.isima.fr/limos/>

Abstract. *Formal Concept Analysis "FCA" is a data analysis method which enables to discover hidden knowledge existing in data. A kind of hidden knowledge extracted from data is association rules. Different quality measures were reported in the literature to extract only relevant association rules. Given a dataset, the choice of a good quality measure remains a challenging task for a user. Given a quality measures evaluation matrix according to semantic properties, this paper describes how FCA can highlight quality measures with similar behavior in order to help the user during his choice. The aim of this article is the discovery of Interestingness Measures "IM" clusters, able to validate those found due to the hierarchical and partitioning clustering methods (AHC and k-means). Then, based on the theoretical study of sixty one interestingness measures according to nineteen properties, proposed in a recent study, FCA describes several groups of measures.*

Key words: Formal Concept Analysis, Interestingness Measure, Properties, Clustering Techniques

1 Introduction

With the development of storing data techniques in modern societies, the user is concerned with the management of large volumes of database. Thus, arises the problem of extracting knowledge from data (KDD, Knowledge Discovering in Databases) introduced by Piatetsky Shapiro [22], where the exponential number of generated rules is an impediment to the user who is facing several interesting and uninteresting rules. Given the problem of finding valid association rules, two objective interestingness measures are used, which are *Support* and *Confidence*. However, these measures are not sufficient to extract only the really interesting knowledge and was challenged in many studies such as [23].

Further to the weakness of the *support-confidence* approach, several interestingness measures [24], [16], [10], [25] have been proposed in the literature to judge the relevance of the extracted rules. However, their large number causes the problem of measures selection. Facing the problem of choosing the right

interestingness measure to extract valid association rules, formal studies of interestingness measures were proposed in [10], [20], [16], [12] for understanding the behavior of different measures. Those studies consist of the evaluation of the interestingness measures according to semantic properties [22], [24], [17], [16] judged important to characterize an interesting measure. This assessment will lead to the construction of a matrix that will be the starting point for an interestingness measures clustering. In this article, we focus on Guillaume and al.'s work [12], [13] which is extended to about sixty measures. Indeed, in a first time, they evaluate sixty one objective measures according to nineteen properties and in a second time, they apply an unsupervised classification on those measures, based initially on the most popular clustering methods: the agglomerative hierarchical method *AHC* [26] and the partitioning method *K-means* [19]. This article is then proposed to validate the clustering methods results. This validation remains a challenging task. Different validity indices were proposed in the literature [15] to measure the quality of the generated clusters. However, none of them is a winner and the user is supposed able to find another way proving the clusters.

Based on *FCA*, we check the identified groups of measures according to the methods listed above and to highlight interestingness measures with similar behavior in order to help the user during his choice. Thus, we propose in this paper the discovery of interestingness measures clusters, able to validate those found due to the hierarchical and partitioning clustering methods (*AHC* and *K-means*), or to highlight new insights.

The second section recalls the basics of Formal Concept Analysis. The third section presents the interestingness measures and defines 19 semantic properties. The fourth section combines clustering techniques and *FCA* results to characterize interestingness measures.

2 Basics of Formal Concept Analysis

2.1 Formal Concept Analysis

Formal Concept Analysis [8] is an approach capable of discovering and structuring knowledge. In this subsection, we present some basic definitions [27] needed for this paper.

Definition 1: Given a formal context $K = (G, M, R)$ consisting of a binary relation R between a set of objects G and a set of attributes M (i.e. $R \subseteq G \times M$). The relation $gRm \Leftrightarrow (g, m) \in R$, is read: "*the object g has the attribute m* ".

For a set of objects $O \subseteq G$, we define:

$$O' := \{m \in M \mid \forall g \in O, (g, m) \in R\};$$

and for a set of properties $A \subseteq M$, we define:

$$A' := \{g \in G \mid \forall m \in A, (g, m) \in R\}.$$

Definition 2: The pair (O, A) is called a formal concept of the context (G, M, R) by Wille [27], with $O \subseteq G$, $A \subseteq M$, $O' = A$ and $A' = O$. The set O is called the extent of the concept (O, A) and A is called its intent. Formally, let L be the set of concepts of (G, M, R) and \leq , the partial order defined between the concepts by: $(O1, A1) \leq (O2, A2) \Leftrightarrow A1 \subseteq A2 \Leftrightarrow O2 \subseteq O1$.

The pair (L, \leq) is called the lattice of context (G, M, R) , which can be graphically represented by the Hasse diagram, which is a downward graph for the understanding of conceptual relationships in data.

2.2 Association rules

In the following, we describe association rules in terms of Formal Concept Analysis, formulated by [21], [28].

Definition 3: An association rule is an implication of the form $X \rightarrow Y$ consisting of two subsets of attributes $X, Y \subseteq M$, called respectively, the antecedent and the consequent of the rule, where $X \cap Y = \emptyset$. We can say that X implies Y if any object with the attributes in X also has the attributes in Y . The support of the rule $X \rightarrow Y$ is $\text{supp}(X \rightarrow Y) := \frac{|g(X \cup Y)|}{|G|}$ (where $|G|$ is the cardinality of G) and its confidence is $\text{conf}(X \rightarrow Y) := \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$.

If the confidence of the rule is equal to 1 ($\text{conf}(X \rightarrow Y) = 1$), then the rule r is called an exact association rule or an implication, otherwise r is called an approximate association rule.

An association rule is valid if and only if its support and confidence are respectively superior to the minimal thresholds, minsupp and minconf . So, for mining association rules [1], there are several algorithms in the literature introduced in [2] based on the two most used interestingness measures, the *support* and the *confidence*. Unfortunately, the *Support-Confidence* approach identifies some weaknesses remaining insufficient for measuring the desired quality.

To address this problem, several measures have been proposed in the literature to eliminate some uninteresting rules and many criteria have been defined to design a good interestingness measure.

3 Interestingness Measures and semantic properties

In the following, we present interestingness measures reported in the literature to extract only relevant association rules and recall nineteen semantic properties necessary to evaluate the measures.

3.1 Interestingness Measures

To identify interesting association rules and to enable the user to focus on what is interesting for him; about sixty interestingness measures [14], [24], [7] exist

in the literature. All of them are defined with at least one of the following parameters: $p(XY)$, $p(\bar{X}Y)$, $p(X\bar{Y})$ and $p(\bar{X}\bar{Y})$, where $p(XY) = \frac{n_{XY}}{n}$ represents the probability of XY , and \bar{X} is the negation of X . However, regarding their high number, the choice of the user becomes very awkward. To help him during his choice of the appropriate measure able to extract the "best rules" from very large databases [2], [4], [3] and responding to his needs, several studies [24], [16], [18], [9] were reported in the literature. The theoretical studies consist of the proposition of semantic properties in order to characterize and evaluate the interestingness measures. The experimental studies classify the measures according to their experimental behavior. The main goal of researchers in the domain is then the user assistance to choose the best interestingness measure, able to respond to his needs. For that, formal properties have been developed [22], [17], [24], [10], [5] in order to evaluate the interestingness measures and to help users understanding their behavior. In the following, we present nineteen semantic properties reported in the literature.

3.2 Semantic properties of measures

This subsection lists 19 of the 21 desired properties for an interestingness measure m , formalized in [12]. We ignore two of them: *P1 (Intelligibility or comprehensibility of the measure)* and *P2 (Easiness to fix a threshold to the rule)* because we think they are very subjective and depend on the user beliefs.

Property 3 : Asymmetric measure. [24], [16]

As the antecedent and the consequent of the rule $X \rightarrow Y$ play different roles, it is desirable for measures assessing differently the rules $X \rightarrow Y$ and $Y \rightarrow X$.

Property 4 : Asymmetric measure in the sense of the conclusion negation [16], [24]

This property takes into account the symmetry in the sense of the conclusion negation. The measure must be able to distinguish between $X \rightarrow Y$ and $X \rightarrow \bar{Y}$.

Property 5 : Measure assessing in the same way $X \rightarrow Y$ and $\bar{Y} \rightarrow \bar{X}$ in the logical implication case. [16]

It is preferable to evaluate in the same way $X \rightarrow Y$ and $\bar{Y} \rightarrow \bar{X}$ in the logical implication case. The evaluation of $\bar{Y} \rightarrow \bar{X}$ reaffirms the implicative relationship of X on Y .

Property 6 : Measure increasing function the number of examples or decreasing function the number of counter-examples. [22], [16]

It is recognized that less a rule has counter-examples, more it is interesting. Therefore, the assessment of the rule's interest can be measured positively according to the high number of examples n_{XY} of the rule, or to the low number of counter-examples $n_{X\bar{Y}}$.

Property 7 : Measure increasing function the data size [10], [24]

As the rule's interest increases according to the data size n , it is interesting to visualize the measure's reaction to the dilatation of the data set. However, if the measure increases with the data size n and provides rule-values close to its maximum value; it would lose its discrimination potential.

Property 8 : Measure decreasing function the consequent/antecedent size. [16], [22]

Given n_X , n_{XY} and $n_{X\bar{Y}}$ fixed values, it is interesting to associate the interest of the rule to the size of Y : if n_Y increases, the measure should decrease. Therefore, the consequent scarcity begets the interest of the premise existence X .

Property 9 : Fixed value a in the independence case. [16], [22]

A good interestingness measure should have a fixed value a in the independence case. We say that X and Y are independent when the prior realization of X does not change the appearance probability of Y and also when the prior realization of Y does not change the appearance probability of X . In this case, the rules $X \rightarrow Y$ and $Y \rightarrow X$ are not relevant since they provide no information even if the confidence has a high value.

Property 10 : Fixed value b in the logical implication case. [16]

The logical rule is a reference situation marking the absence of counter-examples. In fact, a good quality measure must have a fixed value b in the logical implication case (when $p(Y/X) = 1$ or when $X \subseteq Y$).

Property 11 : Fixed value c in the equilibrium case. [6]

In the situation of equilibrium, when the number of examples is equal to the number of counter examples, a good interestingness measure should have a fixed value c .

Property 12 : Identified values in the attraction case between X and Y . [22]

In the attraction case between the antecedent and the consequent of the rule (i.e when $p(XY) > p(X)p(Y)$), it is desirable that an interestingness measure be able to identify the rule-values of the attraction zone.

The value indicated in *property 12* corresponds to the fixed value in the independence case. From this remark, we deduce that if *property 9* is not verified, then *property 12* won't too: if $P9(m) = 0$ then $P12(m) = 0$. This remark is valid to *property 13* too.

Property 13 : Identified values in the repulsion case between X and Y . [22]

In the repulsion case between the antecedent and the consequent of the rule (i.e when $p(XY) < p(X)p(Y)$), it is desirable that an interestingness measure be able to identify the rule-values of the repulsive zone.

Property 14 : Tolerance to the first counter-examples. [16], [25]

To the appearance of the first counter-examples, some authors [11] suggest a constraint on the shape of the curve of an interestingness measure. Some of them suggest the desire to have a slight decrease in the neighborhood of logical implication, rather than rapid or linear decrease. In fact, a nonlinear measure is sensitive to the appearance of counter-examples and then to the noise [16], which reflects the fact that the user may tolerate or not the appearance of few counter-examples without significant loss of rule interest. Measures are then classified between convex, linear or concave shape.

Property 15 : Invariance in case of expansion of certain quantities. [24]

The interestingness measure must be invariant according to the dilation of certain quantities: 1) when we multiply n_{XY} and $n_{X\bar{Y}}$ by a positive constant K_1 and $n_{\bar{X}Y}$ and $n_{\bar{X}\bar{Y}}$ by a positive constant K_2 ;

2) when we multiply $n_{X\bar{Y}}$ and $n_{\bar{X}\bar{Y}}$ by a positive constant K_1 and n_{XY} and $n_{\bar{X}Y}$ by a positive constant K_2 .

Property 16 : Desired relationship between $X \rightarrow Y$ and $\bar{X} \rightarrow Y$ rules. [24]

An interestingness measure m must be able to distinguish between $X \rightarrow Y$ and $\bar{X} \rightarrow Y$. It must verify the following relation: $m(X \rightarrow Y) = -m(\bar{X} \rightarrow Y)$.

Property 17: Desired relationship between $X \rightarrow Y$ and $X \rightarrow \bar{Y}$ antinomic rules. [24]

An interestingness measure m must be able to distinguish between $X \rightarrow Y$ and $X \rightarrow \bar{Y}$. It must verify the following relation: $m(X \rightarrow Y) = -m(X \rightarrow \bar{Y})$.

We have the following relationship with *property 4*: if $P_4(m) = 0$ then $P_{17}(m) = 0$.

Property 18: Desired Relationship between $X \rightarrow Y$ and $\bar{X} \rightarrow \bar{Y}$ rules. [24]

An interestingness measure m must be able to distinguish between $X \rightarrow Y$ and $\bar{X} \rightarrow \bar{Y}$. It must verify the following relation: $m(X \rightarrow Y) = m(\bar{X} \rightarrow \bar{Y})$.

Property 19: Antecedent size is fixed or random. [16]

The premise size is uncertain when the measure is based on a probabilistic model.

Property 20: Descriptive or statistical measure. [16]

A measure is descriptive when its value is invariant in the case of data dilation, when all the numbers are multiplied by the same factor k . Otherwise, it is statistical.

Property 21: Discriminant measure. [16]

A discriminant measure allows us to discern the rules, even when the data set is large.

In the following, we propose a validation of the clustering methods results based on the FCA technique.

4 Combining Clustering Techniques and FCA

Based on Guillaume and al.'s work [12], [13], which consists of the evaluation of 61 interestingness measures according to 19 properties and the construction of a matrix (*61 objects \times 19 attributes*), we form a formal context. However, to obtain it, we choose only the preferred properties, those validated by the measures and therefore considering only the number "1" in the matrix. Except *property 14* (*Tolerance to the first counter-examples*), on which we applied a complete disjunctive coding. We obtained the three following cases for *P14*: *P14.1* (concave measures), *P14.2* (linear measures) and *P14.3* (convex measures) and we keep

only the $P14.1$ in our formal context, since it's the value often recommended by expert for quality measure.

4.1 Clustering results with *CAH* and *K-means*.

We recall firstly the nine groups of measures described in [13] and obtained respectively with *AHC* and *K-means* techniques. There are six measures which belong to different clusters considering the two clustering techniques: $\{Gini, mutual\ information, fukuda, informational\ gain, interest, recall\}$. These measures are not considered yet.

The results of the clustering techniques listed below are realized using Matlab software, and based on the Euclidean distance between pairs of measures and on Ward distance for the aggregation step.

- $C1$: $\{Goodman, Variation\ support, Pearl\}$;
- $C2$: $\{Implication\ index, Dependency, Prevalence, Coverage, J-measure, Gray\ and\ Orlowska's\ Dependency\}$;
- $C3$: $\{Sebag, Least\ contradiction, Descriptive-confirm, Ganascia, Laplace, Confidence, Examples\ and\ counter-examples\}$;
- $C4$: $\{Discriminant\ Probabilistic\ index, Kulczynski, F-measure, Jaccard, Cosine, Support\}$;
- $C5$: $\{Negative\ Reliability, causal\ confidence, causal\ confirmed-confidence, Specificity, Leverage, Putative\ Causal\ dependency, VT100, precision, causal\ confirm\}$;
- $C6$: $\{IIE, IIER, II, likelihood\ link\ index, IPEE, IP3E\}$;
- $C7$: $\{Yule's\ Y, Yule's\ Q, Mgk, Zhang\}$;
- $C8$: $\{Collective\ strength, Cohen, Piatetsky, Correlation, Novelty, Odds\ ratio\}$;
- $C9$: $\{One\ way\ support, Two\ way\ support, Relative\ risk, Loevinger, Conviction, Pavillon, Bayes\ factor, Klosgen\}$.

4.2 Validation process with Formal Concept Analysis.

Using ConExp platform¹, which is a java-based platform developed to build concept lattices, enabling us to have a good visualization of the concept lattices and interpreting the relationships between different concepts. We apply then the formal context (*61 objects and 19 attributes*) on ConExp to be able to visualize the concepts and the groups of measures. We note that the objects of our context are the interestingness measures through which the class hierarchy is accessed. The attributes are the properties of the interestingness measures.

The whole concept lattice holds 338 concepts where special groups of measures can be discovered. For example, if we want to know the set of interestingness measures validating the $P3$ property, we simply select the concept to which $P3$ is attached and all the asymmetric measures will be highlighted.

¹ <http://conexp.sourceforge.net/>

Moreover, the lattice reveals different groups of measures sharing the same properties. In the following, we try to validate the clustering methods results (*AHC* and *K-means*) using the lattice.

To validate the clusters, we focus on the visualization of the concept lattice: given a cluster C_k , for each element in this cluster, we search for similar objects in the lattice and we select all the nodes to which they are attached. By visualizing the selected nodes, we try to find if there exists one concept able to group all the elements of C_k . If we find this central concept, we can argue the cluster obtained by the clustering methods; else many questions arise like should we keep this cluster? or should we split or merge it?

It is always possible to find a formal concept that may contain the elements in the cluster. However those elements may be distant in the sense that they may possess different attributes (or properties). Visualizing the neighborhood of a concept in the lattices highlights close concepts and therefore close objects or attributes.

4.3 Validation of interestingness measures clusters.

The verification of the clustering techniques results reveals the presence of:

- **clusters easily validated:** $C1$, $C4$ and $C8$;
- **clusters hardly validated:** $C3$, $C6$, $C7$ and $C9$;
- **clusters may be questionable:** $C2$ and $C5$.

In fact, the first group presents the clusters which are easily validated through the lattice. The visualization of its elements highlights their closeness. The second group is the one to which the clusters hardly validated belong i.e., clusters validated after intense visualization, especially when we remark the shift of additional measures to them. The last group presents the clusters that may be questionable. We find difficulty to discern some clusters because we are not able to find all its elements in the same group and therefore we haven't any interpretation for this case, even after the clustering realized by *AHC* and *K-means*.

Clusters easily validated: To check $C1$, we remark that the selection of the concept to which *Variation support* measure is attached reveals its presence in the lattice and the closeness of its measures to each others. The shift of the following measures $\{Odds\ ratio, collective\ strength, Cohen, Yule's\ Y, Yule's\ Q, novelty, correlation, piatetsky\}$ to this group is clear in the lattice and we explain it by the fact that all the highlighted measures verify those properties $\{P5, P18\ and\ P21\}$. The visualization of the neighborhood of the selected concept explains the non-membership of those additional measures to $C1$. In fact, the nearest measures are $\{Odds\ ratio, collective\ strength, Cohen\}$ which are closer to their own group than to $C1$. Otherwise, we can validate the $C1$ cluster if its elements are directly linked to each others.

$C4$ cluster is the easiest one to verify according to the lattice, it is sufficient to select the concept which acts as the center of the cluster, the one to which

Discriminant Probabilistic index is attached, to obtain this cluster. Moreover, the lattice reveals two stable subgroups: $\{Kulczynski, Jaccard\}$ and $\{Czekanowski-Dice, Cosine\}$.

The following cluster to check is *C8* and the lattice reveals a very good results. All the measures belonging to this cluster are very close to each others (direct link) and are always found in the same group when we select multiple nodes to which *C8* measures are attached. Moreover, the lattice shows that *Precision* or *Accuracy* is in a direct link with *Odds ratio* and it is explained by the fact that they share $\{P_4, P_5, P_6, P_7, P_8, P_{21}\}$. Then here, the question arises if we have to merge the second subgroup of *C5*, $\{VT100, Precision\}$ to *C8* cluster. However, this don't prevent us to validate the common behavior of all *C8* measures. The visualization of the lattice reveals also the neighborhood of *C8* cluster to *C1*.

Clusters hardly validated: If we continue with *C3* cluster, we note that when we select the concept to which *Least contradiction* is attached, we find all the elements of *C3*, except *Examples and counter-examples index* that we are not able to show with the highlighted measures. However, when we move up by one level in the lattice, we select the *P11* node which is in a direct link with *Least contradiction*, we obtain our group of measures with the presence of $\{IPEE, IP3E, coverage and prevalence\}$. The two last ones are distant from the cluster we are looking for, which argue their non-membership to *C3*. On the contrary, $\{IPEE and IP3E\}$ are very close and we explain it by the fact that they verify *P11* property. To recapitulate, we can validate *C3* cluster after an intense visualization of the lattice.

The *C6* statistical measures behave differently in the lattice and they are divided over three groups: the first one reveals the clustering of $\{IIE, IIER\}$, the second one is $\{II, likelihood link index\}$ and the last one is $\{IPEE, IP3E\}$. Then three groups are resulted from the lattice if we eliminate the distant measures from each of them. However, at a certain level of the lattice and when we select the concept validating $\{P_4, P_5, P_6, P_{14.1}\}$, we find the cluster *C6* with the presence of $\{Zhang, Yule's Q and Example and counter-example\}$, a set of measures which verify also the same previous properties but are closer to their groups than to *C6*. Consequently, we can validate *C6* cluster.

The same method of visualization is also applied to *C7* cluster and we note that it is not possible to find all its elements in the same group. In fact, we find the two following groups: $\{Yule's Y, Yule's Q\}$ and $\{Zhang, Mgc\}$ and we visualize through the lattice that there is no direct link between them. For example, *Figure 1* shows that there is no link between *Mgc* and *Yule's Q*, which explains the fact that we don't find them in the same group. Those groups share an important number of properties $\{P_4, P_5, P_6, P_7, P_9, P_{10}, P_{12}, P_{13}, P_{17}, P_{21}\}$ which prove their similar behavior. But, if we search more and try to visualize the core of the lattice, we find that at a certain level of it, there is one central node able to cluster all the *C7* measures. *Figure 1* illustrate this cluster and explains the relationship between the measures and shows how they are linked to each

others. The proximity between $C7$ measures is also easily readable in the figure and then we can validate $C7$ cluster.

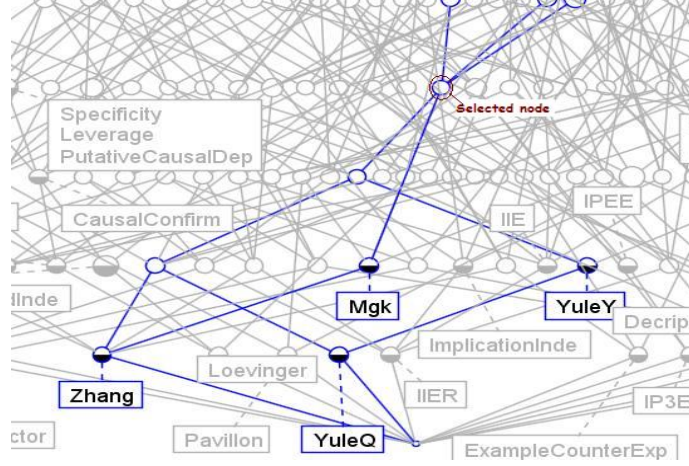


Fig. 1. $C7$ cluster

The last cluster to verify is $C9$, and the concept which acts as center of the group is the one to which "One Way support" is attached (Figure 2). When we select this node, we remark that all the elements of $C9$ are highlighted and we note the presence of some additional measures $\{J\text{-measure, Gray and Orlowska, Dependency, Prevalence and Coverage}\}$ which share some properties with $C9$ measures; some of them are distant but others are near like $\{J\text{-measure, Gray and Orlowska's dependency}\}$. The proximity of the latter and their direct link to *Klosgen* don't prove their membership to $C9$ because they are close to their own group $C2$ too. Figure 2 shows how the $C9$ measures are grouped together, then we can validate it.

Questionable clusters: Concerning $C2$ cluster, it is difficult for us to discern it because any of the 6 measures concepts selected reveal a good results. Only the selection of *Coverage* concept show this group with the presence of all the asymmetric measures, because *Coverage* and $P3$ are attached to the same node. Then, the selection of *coverage* node doesn't help us validating $C2$. However, when we select *J-measure* node, we have the following group $\{J\text{-measure, Dependency and Coverage}\}$. When we select *Gray and Orlowska's dependency*, we have $\{Gray and Orlowska's dependency, Prevalence and Coverage\}$ and finally when we select *Implication index*, we have $\{Implication index, Dependency, Prevalence and Coverage\}$. We remark that there is a link between the $C2$ measures according to the three different groups revealed which make this cluster questionable. Consequently, we can't validate $C2$ cluster.

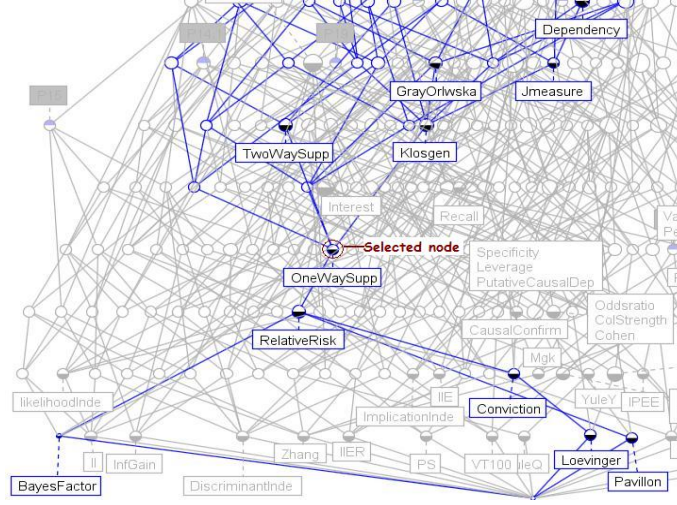


Fig. 2. C9 cluster

The verification of $C5$ measures through the concept lattice shows that we are not able to find them in the same group, at least separated between two groups. On the one hand, we find $\{Negative\ Reliability, causal\ confidence, causal\ confirmed-confidence, causal\ confirm, Specificity, Leverage, Putative\ causal\ dependency\}$ grouped together; on another hand, we have $\{VT100, Precision\}$. As we visualize in the lattice, property $P3$ and $P18$ implies $C5$ measures separation: the first group doesn't validate $P18$, contrary to the second one. The latter doesn't validate $P3$, contrary to the first group. In addition, we note the presence of external measures to this group, some of them are separated by an important number of concepts and others are very close, like *Loevinger* measure which is directly linked to *Negative Reliability*. Then we can say that *Loevinger* shares some properties with $C5$ measures and that it is near to this group of measures even if it belongs to $C9$. Two stable subgroups are revealed from the lattice: $\{Negative\ Reliability, causal\ confidence, causal\ confirmed-confidence\}$ and $\{Specificity, Leverage, Putative\ Causal\ dependency\}$. Finally, we remark that all the $C5$ measures validate $\{P4, P6, P7, P21\}$ and that three of those properties are also shared by $C4$, which explains the closeness of $C4$ and $C5$ that the hierarchical diagram presented in [13] argue too. The question which arises here is if we have to split $C5$ cluster into 2 groups.

4.4 FCA as a third party

FCA can also help to determine the cluster of measures that are in different clusters with the two clustering techniques. In our case, six measures $\{recall, informational\ gain, interest, Gini, mutual\ information, fukuda\}$ are in different clusters.

Concerning *recall* measure, the selection of the node to which it is attached shows its grouping with both *C4* and *C5* measures. But in term of closeness, it is nearer to *C5*, as it is in a direct link with *specificity*. *Informational gain* is directly linked to *interest* and the latter is grouped with both *C8* and *C9* measures i.e., the lattice is not able to affect those measures to any of those clusters. The same thing with *{Gini and mutual information}*, no additional information is revealed concerning them if we remark that graphically, they are near to *C1* like to *C2*. Finally, the selection of *Fukuda* measure reveals that it is close to *C6* measures *{IIE, IIER and IP3E}*.

5 Conclusion

In this paper, we show how FCA can help clustering of interestingness measures. In general, we can notice that FCA validates the clustering results (for both *AHC* and *K-means*). In fact, the verification of some clusters, like *C1* or *C4* is clear through the lattice, especially when we find the concept which acts as center of the group. However, some exceptions have been encountered. For instance, we have difficulty to discern the *C2* cluster. In another hand, we note the separation of some clusters into two (or three) groups like *C5*. Our approach can be generalized to other data sets. It is mainly based on the visualization mode, and can be inefficient when dealing with very large clusters. It is not scalable considering the size of a cluster. Thus we are currently working on a formalization of our approach that could allow to run the process in a semi-automatic way.

References

1. Agrawal, R. Imielinski, T. Swami, A.: Mining association rules between sets of items in large databases. Proc. SIGMOD Conf. 207–216 (1993)
2. Agrawal, R. and Srikant, R.: Fast algorithms for mining association rules. Proc. VLDB Conf. 478–499 (1994)
3. Bayardo, R. J.: Efficiently mining long patterns from databases. Proc. SIGMOD Conf. 85–93 (1998)
4. Brin, S. Motwani, R. Ullman, J. D. and Tsur, S.: Dynamic itemset counting and implication rules for market basket data. Proc. SIGMOD Conf. 255–264 (1997)
5. Blanchard, J. Guillet, F. Briand, H. and Gras, R.: Assessing rule with a probabilistic measure of deviation from equilibrium. In Proc. Of 11th International Symposium on Applied Stochastic Models and Data Analysis ASMDA, pages 191–200, Brest, France. ENST. (2005)
6. Blanchard J. Guillet F. Briand H. and Gras R.: IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles. Dans l'Atelier Qualité des Données et des Connaissances. 26–34 (2005a)
7. Feno D.J.: Mesures de qualité des règles d'association: normalisation et caractérisation des bases, PhD thesis, Université de La Réunion. (2007)
8. Ganter, B. and Wille, R.: Formal Concept Analysis. Mathematical Foundations Edition. Springer. (1999)

9. Geng, L. and Hamilton, H. J.: Interestingness measures for data mining: A Survey. *ACM Computing Surveys*, 38:1–31 (2006)
10. Geng L. and Hamilton H.J.: Choosing the Right Lens : Finding What is Interesting in Data Mining. *Quality Measures in Data Mining*. ISBN 978-3-540-44911-9. 3–24 (2007)
11. Gras, R., Couturier, R., Blanchard, J., Briand, H., Kuntz, P., Peter, P.: Quelques critères pour une mesure de qualité de règles d’association. *RNTI-E-1*. 3–31 (2004)
12. Guillaume, S. Grissa, D. Mephu Nguifo, E.: Propriétés des mesures d’intérêt pour l’extraction des règles. Dans l’Atelier Qualité des Données et des Connaissances, EGC’2010, Hammamet-Tunisie. <http://qdc2010.lri.fr/fr/actes.php>. 15–28 (2010)
13. Guillaume, S. Grissa, D. Mephu Nguifo, E.: Une étude analytique des mesures d’intérêt pour l’extraction des connaissances. Technical Report RR-10-14, LIMOS, ISIMA (2010)
14. Hilderman, R. J. and Hamilton, H. J.: Knowledge Discovery and Measures of Interest, volume 638 of *The International Series in Engineering and Computer Science*. Kluwer. 2, 81 (2001)
15. Kaijun, W. Baijie, W. and Liuqing, P.: CVAP: Validation for Cluster Analyses. *Data Science Journal*. 8:88–93 (2009)
16. Lallich, S. et Teytaud, O.: Evaluation et validation de mesures d’intérêt des règles d’association. *RNTI-E-1*, numéro spécial. 193–217 (2004)
17. Lenca, P, Meyer, P. Picouet, P. Vaillant, B. and Lallich, S.: Critères d’évaluation des mesures de qualité en ecd. *Revue des Nouvelles Technologies de l’Information (Entreposage et Fouille de données)*, (1) :123–134. 2, 82, 84, 95 (2003a)
18. Lenca, P. Meyer, P. Vaillant, B. and Lallich, S.: A multicriteria decision aid for interestingness measure selection. Technical Report LUSI-TR-2004-01-EN, Dpt. LUSI, ENST Bretagne (2004) (chap1)
19. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Math. Statist, Prob.*, 1:281–297 (1967)
20. Maddouri, M. Gammoudi, J.: On Semantic Properties of Interestingness Measures for Extracting Rules from Data. Springer Berlin/Heidelberg. 148–158 (2007)
21. Pasquier, N. Bastide, Y. Taouil, R. Lakhal, L.: Pruning closed itemset lattices for association rules. *Proc. 14ièmes Journées Bases de Données Avancées (BDA’98)*, Hammamet, Tunisie. 177–196
22. Piatetsky-Shapiro G.: Discovery, Analysis and Presentation of Strong Rules. In G. Piatetsky-Shapiro & W.J. Frawley, editors. *Knowledge Discovery in Databases*. AAAI Press. 229–248 (1991)
23. Sese, J. and Morishita, S.: Answering the most correlated n association rules efficiently. In *Proceedings of the 6th European Conf on Principles of Data Mining and Knowledge Discovery*. Springer-Verlag. 410–422 (2002)
24. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. *Information Systems*. 29(4): 293–313 (2004)
25. Vaillant, B.: Mesurer la qualité des règles d’association : études formelles et expérimentales. PhD thesis. ENST Bretagne. (2006)
26. Ward J.H.: Hierarchical Grouping to Optimize an Objective Function. *Journal of American Statistical Association*. 64:236–244 (1963)
27. Wille, R.: Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. In *Ordered Sets*. Ed. by I. Rival. Reidel, Dordrecht-Boston. 445–470 (1982)
28. Zaki, M. J. Ogihara, M.: Theoretical Foundations of Association Rules. 3rd SIGMOD’98 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD). Seattle, WA .7:1-7:8. (1998)